多层次融合的学术文本结构功能识别研究*

■ 王佳敏[1,2] 陆伟[1,2] 刘家伟[1,2] 程齐凯[1,2]
[1] 武汉大学信息管理学院 武汉 430072 [2] 武汉大学信息检索与知识挖掘研究所 武汉 430072

摘要：[目的/意义] 学术文本结构功能是对学术文献的结构和章节功能的概括，针对当前研究较少从学术文本多层次结构出发进行融合和传统方法依赖人工经验构建规则或特征的问题，本文在对学术文本层次结构进行解析的基础上，构建了多层次融合的学术文本结构功能识别模型。[方法/过程] 以ScienceDirect数据集为例进行实验，该模型首先通过深度学习方法对不同层次学术文本进行结构功能识别，接着采用投票方法对不同层次和不同模型的识别结果进行融合。[结果/结论] 研究结果表明各层次集成后的整体效果较单一模型均有不同程度提升，综合结果的整体准确率、召回率和F1值分别达到86%、84%和84%，并且深度学习算法在学术文本分类任务中的性能较传统机器学习算法SVM更优，最后对学术文本结构功能错分情况进行了分析，指出本研究潜在的应用领域和下一步的研究方向。

关键词：深度学习 结构功能 多层次融合 学术文本
分类号：G250
DOI：10.13266/j.issn.0252-3116.2019.13.010

1 引言

近年来，随着科学研究工作的进展，科研论文的产量迎来爆发式增长[1]。以微软学术为例，截至2017年3月，其包含的数据记录达到1.68亿条，并且以每个月130万的记录量持续增长[2]。学术论文是科研人员进行学术研究最主要的信息来源，科研人员在获取学术信息的时候，往往是目标和任务驱动，更加关注文章的某一个特定部分，例如方法、结果或者相关研究的综述等[3-4]，不同结构部分对不同学者的重要性和兴趣也是不一样的[5]。在此背景下，学术文本结构功能识别已经成为学术大数据分析与挖掘领域亟需解决的热点问题[6]。

学术文本结构功能是指对学术文献的结构和章节功能的概括，不同的章节对于论文内容的表达具有特定的功能性作用[7]。学术文本的结构比较规范和固定，具有一定的逻辑和层次。一篇学术论文通常由标题、作者、摘要、关键词和章节组成，章节又包括章节标题、段落、图表、公式、引用等内容[5,8]。尽管学术文本开始越来越规范和标准，尤其是以生物学领域为代表的IMRAD结构的采用[9]，但是这种结构化的论文格式并没有被所有的学术文本和研究领域采纳。而越来越多的研究指出，学术文本的结构功能对于信息检索[3]、关键词抽取[10]、引文分析[11-12]等任务的研究具有明显的提升作用。因此，对大规模学术文本实现结构功能自动识别具有重要的研究意义和实际应用价值。

目前对学术文本结构功能识别的研究主要从文档逻辑结构的视角出发，采用基于规则或基于机器学习的方法，对学术文本不同层次的逻辑结构进行识别，如标题识别[4,7]、章节识别[13-14]、段落识别[15]等，在实际应用中取得了一定的效果，但是依然存在两个问题。第一，对学术文本不同结构部分进行单独识别，而没有从文章的整体层次结构出发，融合多层次结构特征进行识别。实际上学术文本不同层次的文本信息包含了不同的特征和语义信息，综合各个部分的信息可以提供更加完整和准确的判断。第二，传统的基于规则或者机器学习的方法需要人工构建规则或者提取特征，结果的好坏严重依赖于人工经验，迁移能力较低。而

* 本文系国家自然科学基金面上项目"面向词汇功能的学术文本语义识别与知识图谱构建研究"（项目编号：71473183）研究成果之一。
作者简介：王佳敏（ORCID：0000-0003-3954-0381），博士研究生；陆伟（ORCID：0000-0002-0929-7416），副院长，教授，博士生导师，通讯作者，E-mail：weilu@whu.edu.cn；刘家伟（ORCID：0000-0002-2774-1509），硕士研究生；程齐凯（ORCID：0000-0003-3904-8901），讲师，博士。
收稿日期：2018-09-13 修回日期：2019-01-25 本文起止页码：95-104 本文责任编辑：杜杏叶

深度学习可以自动完成数据表示和特征提取,通过学习过程提取出不同水平、不同纬度的有效表示,从而提高不同抽象层次上对数据的解释能力[16],具有增量学习、可迁移性强等特点。为了解决上述问题,本研究在对学术文本多层次结构进行解析的基础上,将学术文本正文划分为 5 个结构功能类别,采用深度学习和投票的方法,构建了多层次融合的学术文本结构功能识别模型,在计算机语言学领域学术文本数据集上对模型的有效性进行测试和评价。

2 相关研究

本研究主要关注从文本内容角度出发对学术文本结构功能的分析和识别,目前对该问题的研究主要围绕文档逻辑结构[17]展开,主要研究方法可以分为基于规则的方法和基于机器学习的方法。基于规则的方法主要通过从文档的布局和文本特征角度出发人工构建规则,实现对学术文本结构的划分。如 J. Kim 等人[18]通过对文档布局的分析和光学字符识别(Optical Character Recognition, OCR)结果特征的抽取构建规则,实现对生物医学学术文献中标题、作者、单位和摘要的自动标注。A. Constantin 等人[19]设计了一个基于规则的系统 PDFX,可以把 PDF 格式的学术文本的逻辑结构进行重组,并从标题、作者、正文和参考文献等语义层面对其进行描述。

基于机器学习的方法将学术文本结构识别转化为文本分类问题,采用相应的机器学习方法进行识别。如 M. T. Luong 等人[8]采用条件随机场的方法实现了学术文本标题、作者、摘要、图表、公式等逻辑结构的识别。S. Tuarob 等人[13]采用随机森林、支持向量机和朴素贝叶斯等机器学习方法,通过划分章节边界实现学术文本章节语义层次的自动识别,虽然结果准确度达到了 92.38%,但是只在 227 篇学术文档上进行了实验。黄永等人从章节标题[7]、章节内容[14]和段落内容[15]三个层次,分别采用 CRF、SVM 等方法实现了学术文本结构功能的自动识别,并取得不错的效果。

深度学习作为机器学习的一个分支,近年来在自然语言处理领域取得快速进展。深度学习概念最早由 G. E. Hinton[20]在 2006 年提出,深度学习通过建立、模拟人脑的分层结构来实现对外部输入的数据进行从低级到高级的特征提取,建立起低级特征到高级语义之间复杂的映射关系。R. Salakhutdinov 等人[21]将深度信念网络(Deep Belief Network, DBN)和堆栈自编码网络用于对文档建立索引以便检索。X. Glorot 等人[22]将深度学习方法用于域自适应情感分类问题,从无监督的在线评论和建议中提取有意义的特征表示,实验结果表明用高阶特征表示训练的情感分类器的学习性能明显优于当前的其他方法。M. M. Rahman 等人[4]将循环神经网络(Recurrent Neural Networks, RNN)引入到文档结构深度理解研究中,并取得不错的效果。

综合来看,传统基于规则的方法需要人工构建规则,算法一般针对特定的文档类型,结果的准确性得不到保证。机器学习的方法较基于规则的方法在识别的精度和效率上有一定的提高,但该方法的缺点在于依靠人工经验抽取样本特征,模型学习后获得的是没有层次结构的单层特征。相比传统机器学习人工构建特征的方式,深度学习能够更加高效的自动完成数据表示和特征提取,通过低层的特征组合,形成更加抽象的深层表示类型或特征,从而提高数据的解释能力,近年来在文本分类领域得到了越来越多地应用。

3 多层次融合结构功能识别模型构建

3.1 学术文本多层次结构

学术文本通常具有严谨的逻辑结构和规范的层次,遵循科学研究的一般过程,从提出研究问题、介绍研究方法到结果的讨论和结论,具有不同目的和功能的章节部分组成了一篇完整的学术文章。本文主要针对学术文本正文部分进行结构功能识别,根据学术文本的逻辑结构,结合前人的研究成果[7,23]和本研究所使用的数据集,将学术文本的正文结构功能划分为“引言”“相关研究”“方法”“实验”和“结论”5 个部分。

学术文本的正文由多个章节组成,每一个章节由章节标题(Section header)和章节内容(Section)组成,每一个章节内容又包括数量不等的段落内容(Paragraph),如图 1 样例所示。结构功能识别就是对学术文本中章节的功能和目的的标注,其本质是一种基于学术文本内容的分类问题,根据研究对象和粒度的不同,其自动识别可以分为三个层次。第一,章节标题层次的结构功能识别,即根据学术文本章节标题文本进行章节功能分类;第二,章节内容层次的结构功能识别,即从学术文本章节全部内容出发进行识别,提供更多的文本特征信息;第三,章节段落层次的结构功能识别,首先对章节内所有段落进行文本分类,根据多数投票原则,所有段落中输出类别最多的那个类就是对应的该章节的结构功能。

Grasping the Structure of Journal Articles: Utilizing the Functions of Information Units

Introduction

Section header

Paragraph 1

Journal articles have been the dominant source of information used by science faculty for academic work (Tenopir, King, Edwards, & Wu, 2009). However, people are overloaded with information brought about by the growing amount of research and the publications readily accessible in a digital world. Furthermore, what is useful or interesting to readers is often a certain part of the journal article rather than the article in its entirety (Bishop, 1999; Sandusky & Tenopir, 2008). During the three decades from 1977 to 2005, for a university science faculty member in the United States, the average number of articles read per year almost doubled, while the average time spent per article read was cut down nearly half (Tenopir et al., 2009). To meet the demand as stated above, this study endeavors to find a way to help readers locate and consume the most relevant information within a scholarly journal article.

Paragraph n

Genre has emerged as a new approach to information system design for supporting information seeking and use. Genre can tell about the function of a document with a few words. Genre knowledge can enable the reader to instantly recognize what the document would look like (form), consequently what it means (content), and ultimately what it is used for (purpose). Genre instances can vary in granularity—an individual web page, multiple web pages, a website, or part of a web page can all be considered as comprising a genre (Crowston & Williams, 2000; Rosso, 2008; Shepherd & Watters, 1998). Most existing studies on digital genre have focused on the level of documents such as journal articles (Dillon, 2004), web pages such as web newspapers (Vaughan, 1999) or websites such as academic and corpo-
structure than articles in other domains. But the research methods and results from this research can be generalized to other genres and different disciplines.

Related Literature

The literature pertaining to the context and the theoretical framework of this research is reviewed in the following two sections.

Genre Studies in Information Science

Genre studies in information science can be applied to knowledge organization, web design, and digital communication (Andersen, 2008). The first two threads are closely related to this research, which indicates the functions of the information units and the structural relations between these functional units.

Methods

A functional unit taxonomy was needed to explicate how the individual functional units in IMRD components were related to the different tasks of using information from journal articles. First, a set of functional units were derived from the move analysis models, then tested with sample articles in psychology domain, and finally validated through two online surveys from the psychology journal users. (For details regarding the development of the functional unit taxonomy see Zhang et al., 2010 and Zhang et al., 2011.)

Results

The collected data were analyzed in terms of effectiveness and efficiency from two perspectives: reading outcomes and reading process, and furthermore the task impact on reading effectiveness and efficiency.

Section

Reading Outcomes

To answer the research questions 1.1a and 2.1a regarding effectiveness and efficiency in reading outcomes, the nonparametric Wilcoxon signed-rank tests were conducted on subjective measures (i.e., questions 3–7 on posttask questionnaires) and objective measures (i.e., through the relevant

Discussion

A complete reading process comprises navigation, close reading, comprehension, and information use. The results of testing show that the functional units can be used to support each stage in the reading process to varied extent. Next, the utilization of functional units is discussed for each stage of reading.

Conclusion

Of the few existing studies on document component use, some identified the generic conventions of four article components based on users' conceptions (Dillon, 2004; Vaughan & Dillon, 1998), while others focused on the individual retrieval of logical subdivisions of an article (Bishop, 1999; Sandusky & Tenopir, 2008). Zhang (2011) claims that the real use of document components depends on identifying the functions of the smallest information units within an article and their connections with information use tasks. This research utilizes a novel system that enables a flexible reading environment in which readers may adjust their

图 1　学术文本结构功能示例

3.2　学术文本结构功能识别模型构建

本文以深度学习技术为基础,融合章节标题、章节内容和章节段落三个层次对学术文本结构功能进行识别,整体研究框架如图 2 所示,主要分为两个模块,基于深度学习的结构功能分类和基于投票方法的多层次融合。该模型的优势在于能够将学术文本不同层次的特征集成在一起,提供了从学术文本正文全局出发的结构功能识别,并巧妙的引入集成学习的思想并采用投票的方法对不同层次的识别结果进行融合。相比单一层次的识别,该模型的应用场景更加丰富,识别结果效果更好。

3.2.1　基于深度学习的结构功能分类　基于深度学习的学术文本结构功能分类模块分为 5 个部分:输入层、词嵌入层、特征学习层、Softmax 层和输出层。输入层即分别由章节标题、章节内容和章节段落内容构成的带标签的训练集和测试集,并将待分类文本统一为同等长度。词嵌入层将输入层的文本转化为向量表示,本研究采用 word2vec 工具来生成词向量模型,将每个词用一个 K 维实向量进行表示。用 word2vec 生成的词向量不仅能很好的解决稀疏性的问题,而且可以通过余弦相似度、欧式距离等方法计算词之间的相似度。特征学习层是整个分类模型中最重要的部分,本文分别采用卷积神经网络(Convolutional Neural Network, CNN)、长短时记忆网络(Long Short Term Memory, LSTM)和 CNN + LSTM 模型对已知类别的词向量表示进行学习,得到训练好的模型,从而对测试集文本进行分类。Softmax 层采用 Softmax 作为激活函数,对特征学习层模型传递来的信息进行学习,计算出待分类数据属于各个类别的概率,是深度学习多分类问题最常用的一个归一化函数。输出层将测试集各类别的预测结果及每一条数据归属各个类别的概率进行输

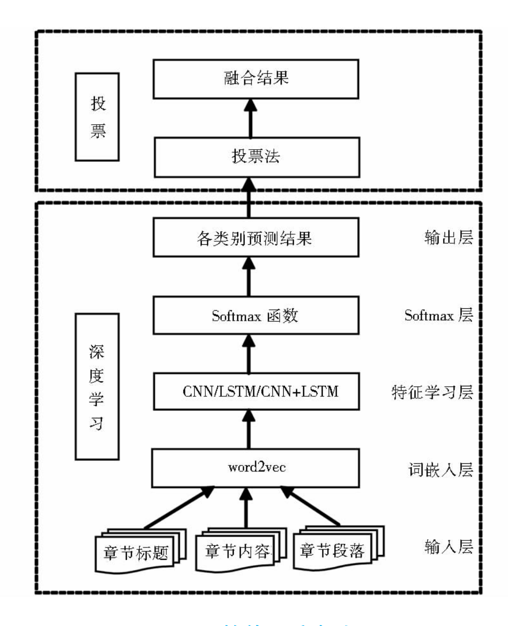


图 2　整体研究框架

出。接下来,分别对 CNN、LSTM 和 CNN + LSTM 模型进行描述。

(1) CNN。本文构建的 CNN 模型结构如图 3 所示,由输入层、卷积层、池化层和输出层组成。

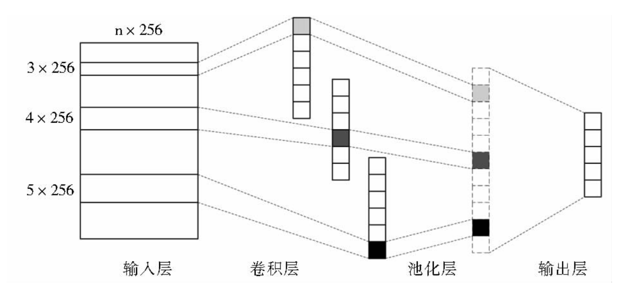


图 3　CNN 结构图

输入层是大小为 $n \times d$ 的学术文本各层次对应的词向量特征矩阵 S,n 表示输入文本的长度,对于长度不足 n 的文本进行补零处理,d 表示词向量维度,本文中 word embedding 长度均为 256。卷积层是对文本特征向量进行高层次特征提取的过程,本文采用 VALID Padding 方式对边界进行处理,步长为 1。对某一卷积核 w,每一步在一个高度为 h 的窗口内进行卷积操作,提取出一个新的特征 c_i:

$$c_i = f(w * S_{i:i+h-1} + b) \qquad 公式(1)$$

其中,f 代表激活函数,本文采用 ReLU 函数作为每个神经元的激活函数,b 代表偏置,h 代表卷积计算中滑动窗口的大小,为了尽可能充分提取出不同粒度大小的局部特征,本文设计了 3、4 和 5 这三种不同大小的卷积核结构组合使用。w 在 $\{S_{1:h}, S_{2:h+1}, \cdots, S_{n-h+1:n}\}$ 这 $n-h+1$ 个窗口进行一轮完整的卷积运算,最终生成特征向量 $C = [c_1, c_2, \cdots c_{n-h+1}]$。

为了获取输入文本中最有用的文本片段,需要对卷积层提取出的特征向量 C 进行最大池化操作(Max Pooling),提取出最大值 $\hat{c} = max(C)$,即寻找对分类结果影响最大的因素。同时,通过池化固定了全连接层的神经元个数,也固定了全连接层输出特征的长度。

最后,在输出层通过全连接的方式将所有得到的局部最优特征连接到最后一层的输出结点,通过 Softmax 函数输出学术文本结构功能的判别结果,并依据训练集中的真实标签,采用反向传播算法对 CNN 中的参数进行梯度更新。

(2) LSTM。LSTM 模型是一种改进的 RNN 模型,针对 RNN 模型存在的梯度消失问题,由 S. Hochreiter 和 J. Schmidhuber[24] 在 1997 年提出。LSTM 用一个记忆单元替换原来 RNN 模型中的隐藏层单元,该记忆单元由记忆细胞(cell)、输入门(input gate)、遗忘门(forget gate)和输出门(output gate)构成[25]。记忆细胞通过状态参数(state)记录信息,并通过相互交互的门单元控制记忆信息值的修改和传递,输入门和输出门负责对参数的输入和输出进行取舍,而遗忘门用来设置选择性遗忘的权重。某时刻 t,LSTM 各结构状态更新公式如下:

$$i_t = sigmoid(W_i * [h_{t-1}, x_t] + b_i) \qquad 公式(2)$$

$$f_t = sigmoid(W_f * [h_{t-1}, x_t] + b_f) \qquad 公式(3)$$

$$o_t = sigmoid(W_o * [h_{t-1}, x_t] + b_o) \qquad 公式(4)$$

$$C_t = f_t * C_{t-1} + i_t * tanh(W_C * [h_{t-1}, x_t] + b_C) \qquad 公式(5)$$

$$h_t = o_t * tanh(C_t) \qquad 公式(6)$$

其中,i_t、f_t、o_t、C_t 分别表示 t 时刻的输入门、遗忘门、输出门和记忆细胞状态,x_t 表示 t 时刻的输入向量,h_t 表示隐藏状态,W_i、W_f、W_o、W_C 和 b_i、b_f、b_o、b_C 分别表示对应的权重矩阵和偏置向量。LSTM 正是通过这种特殊的门结构和记忆单元设置,才能选择哪些信息被遗忘,哪些信息被记住,避免了梯度消失问题,也能学到长周期的信息。

(3) CNN + LSTM。CNN 模型的优点在于能够通过滑动窗口对局部文本进行卷积操作从而提取局部文本特征,缺点在于其对位置不敏感,没有序列刻画的能力。LSTM 模型可以很好的捕捉词汇之间的序列关系,缺点在于其是一个“有偏”模型,次序越靠后的词语越重要。所以本研究尝试对 CNN 和 LSTM 模型进行结合,探究 CNN + LSTM 模型在学术文本结构功能识别中的效果,并与单独使用 CNN 或 LSTM 模型进行对比。CNN + LSTM 的方法,具体来说,就是将 LSTM 的输出作为 CNN 卷积层的输入,将 LSTM 隐藏层的值与 CNN 池化层结果进行结合,最后通过全连接的方式在输出层进行类别输出。

3.2.2　基于投票法的多层次融合　投票法(voting)是集成学习里面针对分类问题的一种结合策略,基本思想是选择所有算法当中输出最多的那个类。相比单个分类算法,集成算法通过多个分类器解决同一个问题,具有更好的泛化能力,结果的质量要高于单个分类算法,在实际应用中取得较好的效果[26]。本文遵循多数投票法的规则,根据公式(7)对识别结果进行融合,即对于每一个章节 x,H,P,S 分别表示该章节对应的章节标题、章节段落和章节内容三个层次的分类结果,R 为经过投票得到的融合结果,即分类结果中得票数最多的类别便为该章节的结构功能。

$$R(x) = Vote(H(x), P(x), S(x)) \qquad 公式(7)$$

4 实验与结果分析

4.1 实验环境

本文中所有的实验均在如表 1 所示的实验环境中完成：

表 1 实验环境及配置

实验环境	环境配置
操作系统	Ubuntu16.04
GPU	NVIDIA GeForce GTX 750 Ti
内存	16G
编程语言	Python3.5
深度学习框架	TensorFlow1.2
word embedding 训练工具	Word2vec

4.2 数据集

本文的实验数据来自 ScienceDirect 数据库 2000 年至 2014 年的计算机语言学领域期刊论文，随机选取其中 4 000 篇学术文献作为本次实验的数据集，共包含 101 种学术期刊，借鉴分层抽样法，将数据集按期刊名分成 101 层，等比例随机抽取其中 3 500 篇文献作为训练集，500 篇文献作为测试集。每一篇文献均包含章节标题、章节内容和章节内所有段落三个层次的文本，其中包含章节标题 21 526 条，章节内容 21 526 条，章节段落 184 433 条。

4.3 评价指标

本文采用准确率（Precision，P）、召回率（Recall，R）和调和平均值（F1）对各个模型的识别结果进行评价，各指标的计算公式如下：

准确率 P = 正确识别的结构功能数/识别出的结构功能数　　公式(8)

召回率 R = 正确识别的结构功能数/实际结构功能数　　公式(9)

调和平均值 $F1 = 2 * P * R / (P + R)$　　公式(10)

整体准确率、召回率和 $F1$ 值为对应各个类别 P、R 和 $F1$ 值的加权算数平均值，作为衡量各模型整体性能的评价指标。

4.4 实验结果及分析

本文分别采用 CNN、LSTM、CNN + LSTM 三种神经网络模型在 Google 开源的 TensorFlow 框架上对章节标题、章节段落和章节内容三个层次的学术文本数据进行结构功能识别的实验。CNN 卷积核窗口高度设置为 3、4、5，LSTM 采用单向模型，2 层隐层，梯度下降优化方法为 Adam，激活函数采用 ReLU，词向量维度为 256，分别对各自的训练数据集迭代学习 200 轮（若模型效果长时间未有提升，则提前终止），其他参数采用单因子变量法实验确立最优参数，通过不断调整超参数训练神经网络模型，直至在训练集上得到最优的实验结果。达到最优实验结果时，其他神经网络模型参数设置如表 2 所示，各模型所需要的训练时间对比如图 4 所示，在每一个模型中从左到右依次对应章节标题、章节段落和章节内容三个层次的结果。

表 2 实验参数设置

模型	章节标题	章节段落	章节内容
CNN	学习率 1e－05 卷积核个数 128 批尺寸 64 丢弃率 0.5	学习率 0.001 卷积核个数 256 批尺寸 16 丢弃率 0.5	学习率 0.001 卷积核个数 128 批尺寸 16 丢弃率 0.5
LSTM	学习率 0.001 隐层节点数 128 批尺寸 64 丢弃率 0.8	学习率 0.001 隐层节点数 128 批尺寸 16 丢弃率 0.8	学习率 0.001 隐层节点数 128 批尺寸 32 丢弃率 0.8
CNN + LSTM	学习率 0.001 卷积核个数 128 隐层节点数 128 批尺寸 64 丢弃率 0.8	学习率 0.001 卷积核个数 256 隐层节点数 128 批尺寸 64 丢弃率 0.8	学习率 0.001 卷积核个数 256 隐层节点数 128 批尺寸 8 丢弃率 0.8

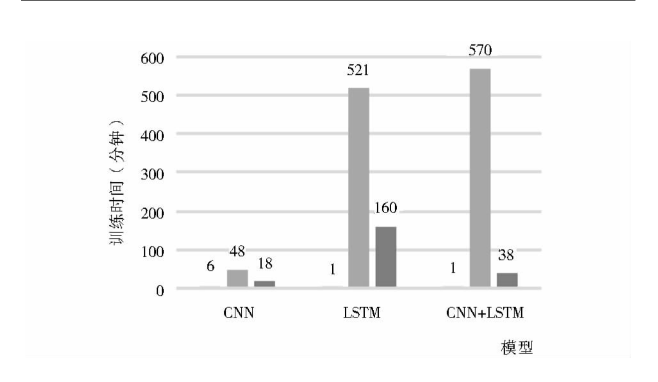


图 4 各模型训练时间对比

从图 4 可以看出，三种深度学习模型在章节段落层次的训练时间最长，章节内容层次次之，章节标题层次时间最短，说明深度学习模型的训练时间与数据量的大小和复杂度成正比。从模型的角度来看，LSTM 模型要比 CNN 模型的训练时间长，这与模型自身的复杂程度相关，而 CNN + LSTM 模型并没有表现出明确的时间增加或减少的现象。

接下来，采用训练好的模型分别对章节标题、章节段落和章节内容层次的测试集进行分类测试，并统计各层次对应的各个类别和整体的准确率、召回率和 F1 值，各层次实验结果分别见表 3、4、5。

从表 3 可以看出，章节标题层次的整体识别准确率均在 85% 以上，三种网络模型效果相差不大，其中

表3 章节标题层次实验结果

章节	CNN			LSTM			CNN + LSTM		
	P	R	F1	P	R	F1	P	R	F1
引言	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
相关研究	0.88	0.63	0.73	0.81	0.66	0.73	0.87	0.61	0.72
方法	0.72	0.87	0.79	0.70	0.86	0.77	0.72	0.84	0.78
实验	0.81	0.77	0.79	0.81	0.73	0.77	0.78	0.79	0.78
结论	0.98	0.91	0.94	0.98	0.91	0.94	0.98	0.90	0.94
整体	0.86	0.85	0.85	0.85	0.84	0.84	0.85	0.85	0.85

表4 章节段落层次实验结果

章节	CNN			LSTM			CNN + LSTM		
	P	R	F1	P	R	F1	P	R	F1
引言	0.80	0.83	0.82	0.83	0.87	0.85	0.83	0.85	0.84
相关研究	0.61	0.15	0.24	0.73	0.15	0.26	0.45	0.23	0.31
方法	0.53	0.76	0.63	0.57	0.82	0.67	0.60	0.68	0.64
实验	0.63	0.74	0.68	0.67	0.74	0.70	0.65	0.78	0.71
结论	0.95	0.43	0.59	0.96	0.53	0.68	0.85	0.60	0.71
整体	0.69	0.65	0.64	0.73	0.70	0.68	0.69	0.69	0.68

表5 章节内容层次实验结果

章节	CNN			LSTM			CNN + LSTM		
	P	R	F1	P	R	F1	P	R	F1
引言	0.86	0.93	0.89	0.92	0.57	0.71	0.84	0.87	0.85
相关研究	0.68	0.24	0.36	0.25	0.03	0.06	0.35	0.26	0.30
方法	0.59	0.79	0.68	0.46	0.50	0.48	0.53	0.72	0.61
实验	0.79	0.66	0.72	0.45	0.42	0.43	0.79	0.60	0.68
结论	0.85	0.84	0.84	0.49	0.88	0.63	0.80	0.79	0.80
整体	0.75	0.74	0.73	0.53	0.52	0.50	0.70	0.68	0.68

效果最好的为CNN模型，其准确率、召回率和F1值分别为86%、85%和85%。在各个结构功能类别的识别结果中，“引言”的识别效果最好，在三种神经网络模型上的准确率、召回率和F1值均达到了100%，“结论”的准确率最高也达到了98%，“相关研究”和“实验”次之。“方法”在三个模型的分类结果中表现较差，准确率最高为72%，通过对语料进行分析，发现该功能部分章节标题表述形式多样，同时由于语料规模的限制，加大了模型对特征的判别难度，导致基于章节标题的“方法”功能识别结果较差，而其他结构功能标题特征的表述较为规范和集中，识别效果较好。

从表4可以看出，章节段落层次的整体识别准确率均在69%以上，其中效果最好的为LSTM模型，其准确率、召回率和F1值分别为73%、70%、68%。其中，CNN + LSTM模型在“方法”类别上的准确率最高，但F1值较LSTM低。在各个结构功能类别的识别结果中，“结论”的准确率最高，达到96%，但召回率仅为53%。“方法”的准确率最低，这一点与章节标题层次的识别结果一致。“相关研究”的召回率较低，说明该功能更容易被错分为其他类别。

从表5可以看出，三种神经网络模型在章节内容层次的整体识别效果相差较大，其中效果最好的为CNN模型，其准确率、召回率和F1值分别为75%、74%、73%，CNN + LSTM模型整体效果次之，LSTM模型整体效果在三者中表现较差。在CNN模型各个结构功能类别的识别结果中，“引言”和“结论”的识别效果最好，其准确率分别达到86%和85%。LSTM模型在“引言”功能上的准确率较CNN高，但F1值较低。“相关研究”的召回率仍然最低，这一点与章节段落层次的识别结果一致。

从各层次识别结果来看，章节标题层次的学术文本结构功能识别效果最好，章节内容层次的识别效果次之，而章节段落层次的识别效果最差。究其原因，章节标题所含的文本较短，且往往直接包含“引言”“相

关研究"“结论”等信息,特征更加明显,模型比较容易学习到有效信息,因此效果最好;而段落和章节内容所包含的文本信息较长,直接以长文本进行学习,加大了模型获取有效特征的难度,效果相对较差。从各模型识别结果来看,三种模型在章节标题层次的识别效果相差最小,在章节内容层次的识别效果相差最大,其中CNN模型在章节标题和章节内容两个层次的整体效果最好,而LSTM模型在章节段落层次的整体效果最好,CNN＋LSTM模型并没有对各层次整体识别效果产生提升。由此也说明,在特征较为明显的短文本分类任务上,CNN、LSTM和CNN＋LSTM模型效果相差不大,随着文本长度的增加,在段落层次LSTM模型表现更好,而当文本长度继续增加到章节层次,CNN模型的表现更好。

接着,本文根据3.2.2小节的投票法对各层次的识别结果进行融合。其中,章节标题投票、章节段落投票和章节内容投票为三种深度学习模型在各个类别上的集成结果,综合投票为三个层次及对应层次下各模型的综合集成结果。投票结果如表6所示:

表6 投票结果

章节	章节标题投票			章节段落投票			章节内容投票			综合投票		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
引言	1.00	1.00	1.00	0.83	0.89	0.86	0.87	0.92	0.90	0.96	0.99	0.98
相关研究	0.89	0.63	0.74	0.76	0.17	0.27	0.70	0.24	0.35	0.94	0.39	0.55
方法	0.72	0.89	0.79	0.59	0.79	0.67	0.59	0.79	0.68	0.68	0.92	0.78
实验	0.82	0.76	0.79	0.67	0.78	0.72	0.79	0.65	0.72	0.87	0.78	0.82
结论	0.98	0.91	0.94	0.95	0.54	0.69	0.81	0.88	0.84	0.97	0.92	0.94
整体	0.87	0.85	0.86	0.74	0.70	0.69	0.75	0.74	0.73	0.86	0.84	0.84

从表6可以看出,经过投票之后,章节标题、章节段落和章节内容层次的整体准确率、召回率和F1值对比投票之前均有小幅度提升或维持不变,说明本文采用多数投票方法对各模型进行集成能够有效提升结构功能识别的整体效果,泛化能力相比单一模型更好。综合投票结果的整体准确率为86%,较章节标题投票结果低1%,较章节段落投票和章节内容投票结果的整体准确率分别提升16.22%和14.67%,说明章节标题是学术文本结构中最有代表性的部分,对整体融合结果的贡献最大。值得注意的是,尽管综合投票结果整体准确率略低于章节标题,但“相关研究”和“实验”两个功能类别的准确率分别比章节标题投票结果提升5.62%和6.10%,说明章节段落和章节内容对这两个功能类别的判断提供了很好的补充,进一步提高了模型的泛化能力。

总体而言,基于投票的融合方法在各层次上均取得不同程度的提升,综合各层次和各模型的融合效果在整体上较章节段落和章节内容的投票结果提升较大,较章节标题层次投票结果稍差。从结构功能类别来看,综合投票在“相关研究”和“实验”两个类别上提升效果较为明显。由此说明,章节标题层次的特征最明显,识别效果最好,一般可直接根据章节标题判断该章节的结构功能。但我们也发现,章节段落和章节内容的词汇特征对结构功能的识别也起到了一定的作用,尤其是在部分类别上展现出较好的优势,对章节标题提供了很好的补充。综合三种层次能够提供更全面的判断,在整体和各类别上均能达到较好的识别效果和泛化能力,在章节标题缺失的情况下通过章节内文本特征对结构功能进行判断也具有较高的可行性。

4.5 对比分析

支持向量机(SVM)是传统机器学习中常用的分类算法,在文本分类中有着非常好的分类效果,也是文献[14]和[15]中采用的分类器。本文采用Python机器学习工具包sklearn进行分类实验,并参照文献[14]和[15]设置,以词汇频次为特征进行结构功能分类。本文通过此实验对比传统机器学习算法与深度学习算法在学术文本结构功能识别中的性能差异。SVM实验结果见表7。

从表7可以看出,相比传统机器学习算法SVM,本文所采用的深度学习算法在章节标题、章节段落和章节内容层次的最优准确率较前者分别提高7.50%、25.86%和20.97%,说明深度学习算法在学术文本结构功能识别任务中的性能要优于传统机器学习算法。SVM作为典型的小样本学习方法,对分类类别规则和特征提取的依赖较大,尤其是在学术文本这种语料相似度较高的文本中,各类别浅层特征不明显,导致支持向量的数量较少,影响了SVM的分类性能。相对来说,深度神经网络能够有效利用句子之间和字与字之间的特征,在学术文本多分类任务中的优势更大。

表 7 SVM 实验结果

章节	章节标题			章节段落			章节内容		
	P	R	F1	P	R	F1	P	R	F1
引言	1.00	1.00	1.00	0.57	0.49	0.52	0.73	0.71	0.72
相关研究	1.00	0.20	0.33	0.64	0.01	0.02	0.38	0.21	0.27
方法	0.70	0.82	0.75	0.49	0.80	0.61	0.58	0.48	0.52
实验	0.57	0.80	0.67	0.61	0.50	0.55	0.56	0.72	0.63
结论	0.97	0.55	0.70	0.61	0.10	0.18	0.76	0.80	0.78
整体	0.80	0.74	0.73	0.58	0.38	0.38	0.62	0.63	0.62

4.6 错误分析

为了进一步发现实验结果中的错误分类情况，本文以综合投票实验为分析对象，输出其分类统计结果，如表 8 所示，其中行代表每一种结构功能被划分为各类别的比例，列代表各结构功能被划分为该类别的比例。

表 8 综合投票实验结果错分表

章节	引言	相关研究	方法	实验	结论	合计
引言	99.00%	0	0.80%	0.20%	0	100%
相关研究	5.58%	39.06%	53.22%	2.14%	0	100%
方法	0.74%	0.59%	91.59%	6.64%	0.44%	100%
实验	0.27%	0.27%	20.42%	77.57%	1.47%	100%
结论	0	0	1.00%	7.42%	91.58%	100%
合计	105.59%	39.92%	167.03%	93.97%	93.49%	500%

从表 8 可以看出，“相关研究”被错分为“方法”的比例最高，被错分为“引言”的比例次之，而其他类别错分为“相关研究”的比例最低。究其原因，在计算机语言学领域学术期刊中，一方面“相关研究”中对方法、模型的介绍比较多，与“方法”部分的文本相似度较高，加大了神经网络模型学习过程中对这两个结构功能的区分难度；另一方面，部分学术论文的“相关研究”章节并没有单独列出，而是融合进“引言”或“方法”部分，导致该功能类别在语料集中所占比例较低。“方法”错分为“实验”的比例最高，而“实验”错分为“方法”的比例最高，说明这两个类别更容易相互错分，结构功能更为相似，这与文献[14]结论一致。从纵向来看，“方法”类别的比例最高，说明其他结构功能更容易错分为“方法”，其中“相关研究”和“实验”是错分为“方法”比例最高的两个类别，这也说明在计算机语言学领域中，对方法的描述在文章各个部分分布较为广泛，而其中“相关研究”和“实验”部分对方法的描述较多。

根据上述错误分析结果，本文认为可以从两个方面尝试进行改进。第一，在各功能类别中增加能代表该类别的词汇特征信息，即选择每一种结构功能与其他结构功能中具有差异的词汇，从而为模型的学习提供更有代表性的特征；第二，增加实验数据量并平衡各功能类别的数量，神经网络模型往往在大规模的数据上才能发挥出优势。

5 结语

本文创新性的将深度学习方法引入到学术文本结构功能识别研究中，分别采用 CNN、LSTM 和 CNN + LSTM 模型对学术文本章节标题、章节段落和章节内容三个层次的文本进行结构功能识别，在此基础上，采用投票的方法探讨了不同层次和不同模型的融合效果。从各层次识别结果来看，章节标题层次整体效果最优，章节内容层次次之，章节段落层次较差。从模型角度来看，CNN 模型综合表现比 LSTM 模型好，而 CNN 与 LSTM 的结合模型在本研究的分类任务中并没有比单独使用 CNN 或 LSTM 模型的效果好。通过与传统机器学习算法 SVM 对比，深度学习算法在学术文本分类任务中的性能更优。从融合结果来看，各层次投票后的整体效果较投票之前均有不同程度提升，而综合投票结果的整体准确率、召回率和 F1 值分别达到 86%、84% 和 84%。整体来看，本研究提出多层次融合的学术文本结构功能识别模型高效可行，具有实际应用价值和增量学习、迁移学习的能力。

在学术大数据环境下，学术文本的挖掘向细粒度和深层语义理解方向发展，从结构功能角度理解学术文本能够促进相关研究向更深层次发展。例如，从结构功能角度出发对不同章节的词汇功能、词汇语义角色进行分析，提供更细粒度的研究；将文本结构功能与引文功能进行融合能够为引文推荐、知识结构的发现提供新的视角；此外，还可以探索基于学术文本结构功能的学术论文评价，为基于内容的论文评价提供支撑。

在下一步的研究中，本文将从两个方面尝试对模型进行改进。一是深度学习在提取文本特征时候，通过 word embedding 将词语转化为固定长度的向量表

示,实际结果出现部分类别识别度较低和类别之间互相错分的问题。为此,本文拟尝试从各章节中词汇的功能特征和语义特征等角度增加文本分类特征,进一步提高类别之间的区分度,并将探索注意力机制在提高模型对文本的理解能力和对各个结构功能类别的区分能力方面的应用。二是增加语料规模,由于深度学习模型的训练对数据量的要求较高,在大规模的数据集上能发挥较大的优势,本文将结合人工标注和机器标注构建多领域大规模的结构功能数据集,进一步提升模型的性能和泛化能力。

参考文献:

[1] XIA F, WANG W, BEKELE T M, et al. Big scholarly data: a survey[J]. IEEE transactions on Big Data, 2017, 3(1): 18－35.

[2] HUG S E, BRANDLE M P. The coverage of microsoft academic: analyzing the publication output of a university[J]. Scientometrics, 2017, 113(3): 1551－1571.

[3] RIBAUPIERRE H D, FALQUET G. Extracting discourse elements and annotating scientific documents using the sciannotdoc model: a use case in gender documents[J]. International journal on digital libraries, 2017, 18(3): 1－16.

[4] RAHMAN M M, FININ T. Deep understanding of a document's structure[C]// Proceedings of the 3rd IEEE/ACM international conference on Big Data computing, applications and technologies. Shanghai:IEEE/ACM, 2017: 63－73.

[5] ALZAHRANI S, PALADE V, SALIM N, et al. Using structural information and citation evidence to detect significant plagiarism cases in scientific publications[J]. Journal of the American Society for Information Science and Technology, 2012, 63(2): 286－312.

[6] KHAN S, LIU X F, SHAKIL K A, et al. A survey on scholarly data: from big data perspective[J]. Information processing and management, 2017, 53(4): 923－944.

[7] 陆伟, 黄永, 程齐凯. 学术文本的结构功能识别——功能框架及基于章节标题的识别[J]. 情报学报, 2014, 33(9): 979－985.

[8] LUONG M T, NGUYEN T D, KAN M Y. Logical structure recovery in scholarly articles with rich document features[J]. International journal of digital library systems, 2010, 1(4): 1－23.

[9] SOLLACI L B, PEREIRA M G. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey[J]. Journal of the medical library association, 2014, 92(3): 364－367.

[10] 方龙, 李信, 黄永, 等. 学术文本的结构功能识别——在关键词自动抽取中的应用[J]. 情报学报, 2017, 36(6): 599－605.

[11] HU Z G, CHEN C M, LIU Z Y. Where are citations located in the body of scientific articles? a study of the distributions of citation locations[J]. Journal of informetrics, 2013, 7(4): 887－896.

[12] DING Y, LIU X Z, GUO C, et al. The distribution of references across texts: some implications for citation analysis[J]. Journal of informetrics, 2013, 7(3): 583－592.

[13] TUAROB S, MITRA P, GILES C L. A hybrid approach to discover semantic hierarchical sections in scholarly documents[C]// Proceedings of the 13th international conference on document analysis and recognition. Mancy:IAPR, 2015: 1081－1085.

[14] 黄永, 陆伟, 程齐凯. 学术文本的结构功能识别——基于章节内容的识别[J]. 情报学报, 2016, 35(3): 293－300.

[15] 黄永, 陆伟,程齐凯,等. 学术文本的结构功能识别——基于段落的识别[J]. 情报学报, 2016, 35(5): 530－538.

[16] 奚雪峰, 周国栋. 面向自然语言处理的深度学习研究[J]. 自动化学报, 2016, 42(10): 1445－1465.

[17] MAO S, ROSENFELD A, KANUNGO T. Document structure analysis algorithms: a literature survey[J]. Proc spie electronic imaging, 2003(5010): 197－207.

[18] KIM J, LE D X, THOMA G R. Automated labeling in document images[C]// Proceedings of the SPIE conference on document recognition and retrieval VIII. San Jose:SPIE, 2000: 111－122.

[19] CONSTANTIN A, PETTIFER S, VORONKOV A. PDFX: fully-automated PDF-to-XML conversion of scientific literature[C]// Proceedings of the ACM symposium on document engineering. Florence:ACM, 2013: 177－180.

[20] HINTON G E, SALAKHUTDINOY R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504－507.

[21] SALAKUTDINOV R, HINTON G E. Semantic hashing[J]. International journal of approximate reasoning, 2009, 50(7): 969－978.

[22] GLOROT X, BORDES A, BENGIO Y. Domain adaptation for large-scale sentiment classification: a deep learning approach[C]// Proceedings of the 28th international conference on machine learning. Washington: Ominpress, 2011: 513－520.

[23] ZHANG L. Grasping the structure of journal articles: utilizing the functions of information units[J]. Journal of the Association for Information Science and Technology, 2012, 63(3): 469－480.

[24] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735－1780.

[25] GRAVES A. Supervised sequence labelling with recurrent neural networks[D]. München: Technische Universität München, 2008.

[26] 章成志. 基于集成学习的自动标引方法研究[J]. 情报学报, 2010, 29(1): 3－8.

作者贡献说明:

王佳敏:负责实验分析和数据处理,论文撰写;

陆伟:提出研究思路和框架,论文修改;

刘家伟:负责实验分析,论文修改;

程齐凯:参与论文框架整理,论文修改。

Research on Structure Function Recognition of Academic Text Based on Multi-level Fusion

Wang Jiamin[1,2] Lu Wei[1,2] Liu Jiawei[1,2] Cheng Qikai[1,2]
[1] School of Information Management, Wuhan University, Wuhan 430072
[2] Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan 430072

Abstract: [**Purpose/significance**] The structure function of the academic text refers to the summarization of academic text structure and section function. While few of existed studies pay attention to the fusion of multi-level structure of academic text, and the traditional methods usually rely on artificial experience to build rules or features. After the analysis of the multi-level structure of academic text, we construct a structure function recognition model based on multi-level fusion. [**Method/process**] We use the academic text dataset from ScienceDirect for experiment. First, we apply deep learning algorithms to identify the structure function of academic text at different level. Then we employ the voting method to fuse the results from different levels and models. [**Result/conclusion**] The results show that the performance improved to varying degrees after fusion. The precision, recall and F1 value of the combined results reached 86%, 84% and 84%, respectively. Compared with the traditional machine learning algorithm SVM, the deep learning algorithm has better performance in the task of academic text classification. Finally, we analyze the misclassification of the structure function of academic text and point out the potential application fields and future research directions.

Keywords: deep learning structure function multi-level fusion academic text